

LEARNING A POSE LEXICON FOR SEMANTIC ACTION RECOGNITION

Lijuan Zhou, Wanqing Li, Philip Ogunbona

School of Computing and Information Technology
University of Wollongong, Keiraville, NSW 2522, Australia
lz683@uowmail.edu.au, wanqing@uow.edu.au, philipo@uow.edu.au

ABSTRACT

This paper presents a novel method for learning a pose lexicon comprising semantic poses defined by textual instructions and their associated visual poses defined by visual features. The proposed method simultaneously takes two input streams, semantic poses and visual pose candidates, and statistically learns a mapping between them to construct the lexicon. With the learned lexicon, action recognition can be cast as the problem of finding the maximum translation probability of a sequence of semantic poses given a stream of visual pose candidates. Experiments evaluating pre-trained and zero-shot action recognition conducted on MSRC-12 gesture and WorkoutSu-10 exercise datasets were used to verify the efficacy of the proposed method.

Index Terms— Lexicon, semantic pose, visual pose, action recognition.

1. INTRODUCTION

Human action recognition is currently one of the most active research topics in multimedia content analysis. Most recognition models are typically constructed from low to middle level visual spatio-temporal features and directly associated with class labels [1, 2, 3, 4, 5]. In particular, many methods [1, 6, 7, 8] have been developed based on the concept that an action can be well represented by a sequence of key or salient poses and these salient poses can be identified through visual features alone. However, these salient poses often do not necessarily possess semantic significance thus leading to the so-called semantic gap. We refer to the salient poses defined using visual features as “visual poses”.

The textual instruction of how a simple or elementary action should be performed is often expressed using Talmy’s typology of motion [9]. In this typology a motion event is one in which an entity experiences a change in location with respect to another object and such event can be described by four semantic elements including *Motion (M)*, *Figure (F)*, *Ground (G)* and *Path (P)*, where *M* refers to the movement of an entity *F* that changes its location relative to the reference object *G* along the trajectory *P* (including a start and an end). For a complex action or an action that involves many body parts,

a sequence of basic textual instructions will suffice. Each instruction, describing a part of the action, will be made up of the four elements. A simplification of the trajectory *P* for human actions entails retaining the starting P_s and end P_e , status or configuration of the body parts, and ignoring intermediate parts of the trajectory. We refer to both P_s and P_e as “semantic poses”. Hence, an action can be described by a sequence of semantic poses if *F* is defined as the whole human body. Alternatively, if *F* refers to a body part, multiple sequences of semantic poses can be used. Semantic poses can be obtained by parsing the textual instructions.

This paper proposes a method to construct a pose lexicon comprising a set of semantic poses and the corresponding visual poses, by learning a mapping between them. It is assumed that for each action there is a textual instruction from which a set of semantic poses can be extracted through natural language parsing [10] and that for each action sample a sequence of visual pose candidates can be extracted. The mapping task is formulated as a problem of machine translation. With the learned lexicon, action recognition can be considered as a problem of finding the maximum posterior probability of a given sequence of visual pose candidates being generated from a given sequence of semantic poses. This is equivalent to determining how likely the given sequence of visual pose candidates follow a sequence of semantic poses. Such a lexicon bridges the gap between the semantics and visual features and offers a number of advantages including text-based action retrieval and summarization, recognition of actions with small or even zero training samples (also known as zero-shot recognition), and easy growth of semantic poses for new action recognition since poses in the lexicon are sharable by many actions.

The rest of this paper is organized as follows. Section 2 provides a review of previous work related to semantic action recognition. The proposed method for learning pose lexicon and action classification is developed and formulated in Section 3. In Section 4, experiments are presented to demonstrate the effectiveness of the proposed method in recognition tasks using MSRC-12 Kinect gesture [11], WorkoutSU-10 exercise [12] datasets and novel actions extracted from the two datasets. Finally, the paper is concluded with remarks in Section 5.

2. RELATED WORK

Despite the good progress made in action recognition over the past decade, few studies have reported methods based on semantic learning. Earlier methods bridged the semantic gap using mid-level features (eg. visual keywords) [13]) obtained by quantizing low-level spatio-temporal features which form visual vocabulary. However, mid-level features are not sufficiently robust to obtain good performance on relatively large action dataset. This problem has been addressed by proposing high-level latent semantic features to represent semantically similar mid-level features. Unsupervised methods [14, 15] were previously applied for learning latent semantics based on topic models; example include probabilistic latent semantic analysis [16] and latent Dirichlet allocation (LDA) [17]. Recently, multiple layers model [18] based on LDA was proposed for learning local and global action semantics. The intuitive basis of using mid- and high-level latent semantic features is that frequently co-occurring low-level features are correlated at some conceptual level. It is noteworthy that these two kinds of semantic features have no explicit semantic relationship to the problem; a situation different from the proposed semantic poses.

Apart from learning latent semantics of actions, other approaches focused on defining semantic concepts to describe action or activity related properties. Actions were described by a set of attributes that possess spatial characteristics. Unfortunately, the attributes are not specific enough to allow subjects to recreate the actions [19, 20]. It is also difficult to describe them as there is no a common principle for describing different actions. In our work, textual instructions use a common principle (four semantic elements) for action representation and they also provide unambiguous guideline for performing actions. Activity was represented by basic actions and corresponding participants such as subjects, objects and tools [21, 22]. The method for representing activities ties the object and action together. The work presented in this paper focuses on single actions which do not depend on other objects.

3. PROPOSED METHOD

Visual poses can be extracted from either RGB, depth maps or skeleton data. In this work, we take skeleton data as an example to illustrate the proposed method. Inspired by the translation model for image annotation [23], a translation model from visual poses to semantic poses, namely “visual pose-to-semantic pose translation model” (VS-TM), is proposed for learning a pose lexicon from skeleton data with instructions. Figure 1 illustrates the action recognition framework based on the VS-TM. In the training phase, a pose lexicon is constructed in two main steps. First, a parallel corpus is constructed based on a stream of semantic poses and a stream of visual pose candidates. Second, a mapping between the two streams is learned from the parallel corpus to generate

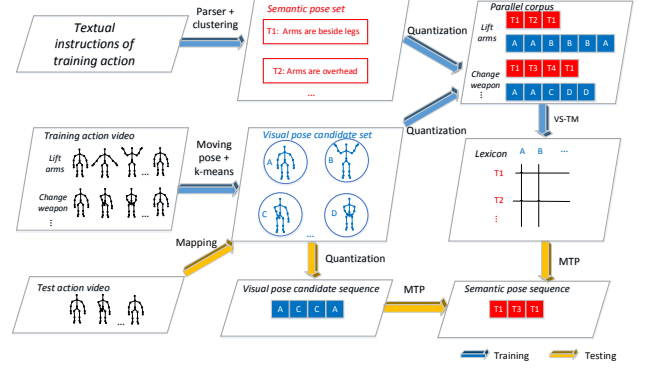


Fig. 1: Framework of the proposed method.

the lexicon for inferring optimal visual poses from the candidates. In the test phase, actions are classified according to the maximum translation probability (MTP) of a semantic pose sequence given a sequence of visual pose candidates.

3.1. Parallel corpus construction

An action instance is represented by two streams; semantic pose and visual pose candidate. The parallel corpus consists of multiple such data streams that have been constructed from each action instance through vector quantization on the sets of semantic pose and visual pose candidate. In the following Sections 3.1.1 and 3.1.2 we focus on how to construct the two sets.

3.1.1. Semantic poses generation

Semantic poses are constructed based on start and end semantic poses P_s and P_e . P_s refers to the configuration of human body in which the action starts and P_e indicates the point at which body parts reach a salient configuration, e.g. maximum extension. P_s and P_e are normally encoded by preposition phrases in the textual instruction. Constituency-based parser (i.e. Berkeley [10]) can be used for extracting P_s and P_e . Note that parsing is not the focus of paper. Suppose an action instance contains G elementary or simple actions. The semantic pose sequence can be written as $\{P_{s1}, P_{e1}, \dots, P_{si}, P_{ei}, \dots, P_{sG}, P_{eG}\}$, where P_{si} and P_{ei} denote semantic poses of the i -th elementary action. Once the semantic poses are extracted, similar semantic poses are replaced by a single symbol and the semantic pose set is easily constructed.

3.1.2. Visual pose candidates generation

Key frames are firstly extracted from each action instance to visually represent the start and end semantic poses. Visual pose candidates are further generated through clustering key frames of all action instance. In this paper, we consider key frames as frames when body parts reach maximum or minimum extension. If we consider skeleton joints as observa-

tions, the covariance matrix of joint positions at an instant time captures the stretch of body parts or distribution of joints in space at this instant time. Hence, the covariance of joint positions is applied for extracting key frames.

Given an action instance containing F frames, F feature vectors are generated to represent this instance using a moving pose descriptor [24]. A covariance is calculated from each feature vector, resulting in a 3×3 matrix. Suppose Σ^f denote the covariance matrix at frame f ($f \in \{1, \dots, F\}$). The relationships amongst the joints of the pose can be analysed by performing eigen-decomposition on Σ^f . For each frame we select the largest eigenvalue, denoted by λ^f and thus, produce the sequence $\Lambda = \{\lambda^1, \lambda^2, \dots, \lambda^f, \dots, \lambda^F\}$.

To reduce the impact of noise in skeletons, we smoothen the sequence Λ along the time dimension, with a moving Gaussian filter of window size 5 frames. Then frames, at whose smoothed largest eigenvalues are bigger or smaller than those of neighbouring frames, are extracted as key frames. In particular, frame f is a key frame if the following conditions are met:

$$\begin{aligned} \lambda^f &> \lambda^{f+1} & \text{or} & \lambda^f < \lambda^{f+1} \\ \lambda^f &> \lambda^{f-1} & \text{or} & \lambda^f < \lambda^{f-1}. \end{aligned} \quad (1)$$

The extracted key frames of all action instances are clustered using k -means algorithm and cluster centers are considered as visual pose candidates. The selection of k depends on the size of semantic pose set and the size of semantic pose set is determined by the number of elementary actions. Hence, it is available before learning the lexicon. One semantic pose can be mapped to multiple visual pose candidates when these visual pose candidates are similar. However, one visual pose candidate corresponds to at most one semantic pose. Therefore, k is chosen as equal or larger than the number of semantic poses so that any semantic pose can correspond to at least one visual pose candidate.

3.2. Lexicon Learning

3.2.1. Formulation

Given the parallel corpus which has underlying correspondence between sequences of semantic pose and visual pose candidate, the problem of learning an action lexicon entails determining precise correspondences among the elements of the two sequences. Let the set of visual pose candidates be denoted by $S = \{S_1, S_2, \dots, S_p, \dots\}$ and the semantic pose set by $T = \{T_1, T_2, \dots, T_q, \dots\}$. The task of lexicon construction is converted to finding the most likely visual pose candidate given a semantic pose, based on conditional probability $P(S_p|T_q)$.

The underlying correspondence between the two sequences provides an opportunity to use machine translation to model the problem. The sequence pair encodes the starting and end positions of actions which are discrete units. These discrete units are actually analogous to words in translation

model. This observation makes the particular word-based machine translation framework useful to our problem. The sequence of visual pose candidate is analogous to the source language and semantic pose sequence is similar to the target language. According to the standard word-based translation framework [25], learning a lexicon is converted to the particular problem - translation model. Hence, we develop a translation model from visual poses to semantic poses (VS-TM) based on the parallel corpus to learn a pose lexicon.

We now illustrate the translation model (VS-TM). Let M_n denote the number of visual pose candidates in the n -th action instance. The sequence of visual pose candidate, after quantization, can be written as $s_n = \{s_{n1}, s_{n2}, \dots, s_{nj}, \dots, s_{nM_n}\} (s_{nj} \in S)$. Similarly, if L_n represents the number of semantic poses in the n -th action instance, then the semantic pose sequence of the action instance can be written as $t_n = \{t_{n1}, t_{n2}, \dots, t_{ni}, \dots, t_{nL_n}\} (t_{ni} \in T)$. VS-TM finds the most likely sequence of visual pose candidate for each semantic pose sequence through the conditional probability $P(s_n|t_n)$.

In the word-based translation model, the conditional probability of two sequences is converted to the conditional probability of elements of the sequences. However, we do not know the correspondence between individual elements of the sequence pair. If we introduce a hidden variable a_n which determines the alignment of s_n and t_n , the alignment of N sequence pairs form a set which can be written as $a = \{a_1, a_2, \dots, a_n, \dots, a_N\}$. Based on a_n , the translation of sequence pair is accomplished through summing conditional probabilities of all possible alignments. Hence, we learn VS-TM through element-to-element alignment models. $P(s_n|t_n)$ is calculated using

$$P(s_n|t_n) = \sum_{a_n} P(s_n, a_n|t_n). \quad (2)$$

If each visual pose candidate in the sequence can be aligned to at most one semantic pose, we guarantee that a visual pose candidate corresponds to only one semantic pose. To ensure this constraint, we construct the alignment from visual to semantic pose sentence. The alignment of the n -th instance can be written as $a_n = \{a_{n1}, a_{n2}, \dots, a_{nj}, \dots, a_{nM_n}\} (a_{nj} \in [0, L_n])$, where a_{nj} represents the alignment position of the j -th visual pose candidate. If the j -th visual pose candidate is aligned to the i -th semantic pose, we write, $a_{nj} = i$. $a_{nj} = 0$ refers to the situation in which no semantic pose corresponds to this visual pose candidate; this happens when visual pose candidate is noisy. According to alignment a_n , Equation (2) can be extended through structuring $P(s_n, a_n|t_n)$ without loss of generality by chain rule as follows:

$$\begin{aligned} P(s_n|t_n) &= \sum_{a_n} \prod_{j=1}^{M_n} P(a_{nj}|s_{n1}^{(j-1)}, a_{n1}^{(j-1)}, t_{n1}^{L_n}) \\ &\quad \times P(s_{nj}|s_{n1}^{(j-1)}, a_{n1}^{(j-1)}, t_{n1}^{L_n}), \end{aligned} \quad (3)$$

where the first item determines alignment probability, the second encodes translation probability and $x_{n1}^{nj} = \{x_{n1}, x_{n2}, \dots, x_{nj}\} (x \in (s, t, a))$. Since lexicon acquisition aims to find conditional probability among visual and semantic poses, we further assume that alignment probabilities are equal (i.e. $\frac{1}{L_n+1}$) and s_{nj} depends only on the sequence element at a_{nj} position which is $t_{na_{nj}}$ (equal to t_{ni}). Hence, Equation (3) can be rewritten as

$$P(s_n|t_n) = \sum_{a_{nj}=0}^{L_n} \prod_{j=1}^{M_n} \frac{1}{L_n+1} P(s_{nj}|t_{na_{nj}}), \quad (4)$$

where the translation probability is constrained through $\sum_{S_p} P(S_p|T_q) = 1$ for any T_q .

For N action instances in the training parallel corpus, the proposed model VS-TM aims to maximize the translation probability $P(s|t)$ through

$$P(s|t) = \prod_{n=1}^N \prod_{j=1}^{M_n} \sum_{i=0}^{L_n} \frac{1}{L_n+1} P(s_{nj}|t_{ni}). \quad (5)$$

Here, it is easy to verify that the sum can be interchanged in Equation (4).

3.2.2. Optimization

The expectation maximization (EM) algorithm is invoked for the optimization by mapping the translation probability $P(S_p|T_q)$ to parameter θ and alignment a to the unobserved data. The likelihood function \mathcal{L} is defined as

$$\mathcal{L}_\theta(s, t, a) = \prod_{n=1}^N \prod_{j=1}^{M_n} \sum_{i=0}^{L_n} \frac{1}{L_n+1} P(s_{nj}|t_{ni}). \quad (6)$$

Maximizing likelihood function \mathcal{L} is further extended to seek an unconstrained extremum of auxiliary function

$$h(P, \beta) \equiv \prod_{n=1}^N \prod_{j=1}^{M_n} \sum_{i=0}^{L_n} \frac{1}{L_n+1} P(s_{nj}|t_{ni}) - \sum_{T_q} \beta \left(\sum_{S_p} P(S_p|T_q) - 1 \right) \quad (7)$$

In the E-step, the posterior probability among alignment is calculated by

$$P_\theta(a_{nj}|s_{nj}, t_{ni}) = \frac{P(s_{nj}|t_{ni})}{\sum_{i=0}^{L_n} P(s_{nj}|t_{ni})} \quad (8)$$

In the M-step, parameter θ is updated through

$$P(S_p|T_q) = \beta^{-1} \sum_{n=1}^N \sum_{j=1}^{M_n} \sum_{i=0}^{L_n} P_\theta(a_{nj}|s_{nj}, t_{ni}) \times \delta(S_p, s_{nj}) \delta(T_q, t_{ni}). \quad (9)$$

Here, $\delta(., .)$ is 1 if two elements are equal and 0 otherwise. β normalizes the probabilities.

3.3. Action classification

Once the translation model from visual pose candidates to semantic poses is learned, the task of action classification is converted to finding the most likely semantic pose sequence given a sequence of visual pose candidate. This is the decoding process in machine translation system [25]. Since textual instructions of all action classes are available, we reduce search space to the possible solution space containing instructions of all trained actions.

Let $s^{test} = \{s'_1, \dots, s'_j, \dots, s'_m\} (s'_j \in S)$ denote the sequence of visual pose candidate in a test action instance and $t^{test} = \{t'_1, \dots, t'_i, \dots, t'_l\} (t'_i \in T)$, its semantic pose sequence. The alignment a^{test} is written as $a^{test} = \{a'_1, \dots, a'_j, \dots, a'_m\}$. Given the model parameter θ , action is classified based on $P(s^{test}|t^{test})$ which is calculated through finding the best alignment to avoid summing all possible alignment probability. In particular, it is formulated as

$$\{t^{test}, a^{test}\} = \arg \max_{\{t^{test}, a^{test}\}} \prod_{j=1}^m P_\theta(a'_j|s'_j, t'_i). \quad (10)$$

4. EXPERIMENTS AND RESULTS

4.1. Datasets and experimental setup

Two action datasets, MSRC-12 Kinect gesture [11] and WorkoutSU-10 [12], were used to evaluate our method. MSRC-12 Kinect gesture dataset is collected from 30 subjects performing 12 gestures and contains 594 video sequences of skeletal body part movements. Each sequence contains 9 to 11 instances. In total, there are 6244 instances in this dataset. WorkoutSu-10 dataset is collected from 15 subjects performing 10 fitness exercises. Each exercise has been repeated 10 times by each subject. It contains 510550 frames resulting in 1500 instances and large time span for each instance. The large time spans make our method valuable in key frames extraction.

Textual instructions of actions in the two datasets and linguistics description of extracted semantic poses are manually described which shown in supplementary material¹. However, it is not difficult to automatically obtain textual instructions as they are often used by sports trainers and may be available online. In total, 16 and 15 semantic poses are respectively applied to MSRC-12 and WorkoutSu-10 datasets. The ground truth lexicon can be seen in Figure 2 which illustrates corresponding optimal visual pose of semantic pose. Here, symbols have been used to represent semantic poses.

Experiments were conducted to evaluate the proposed method on two classification problems including trained and untrained actions. In order to compare with the state-of-the-art algorithms, cross-subject evaluation scheme was applied on the instance level of actions. Moreover, the initial value of

¹ Supplementary material is attached at the end of paper.

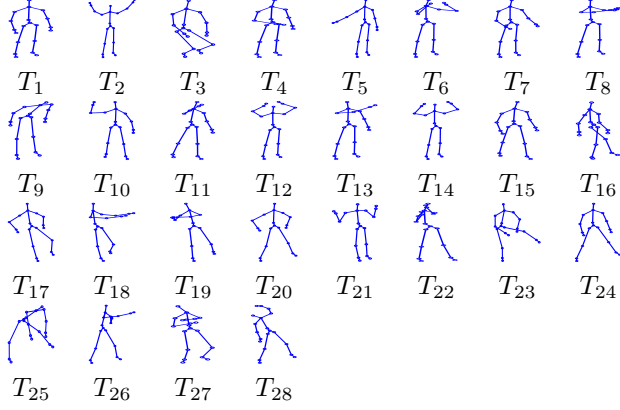


Fig. 2: Ground truth optimal visual poses corresponding to semantic poses.

translation probability $P(S_p|T_q)$ is assigned a uniform probability based on $\sum_{S_p} P(S_p|T_q) = 1$.

4.2. Results

4.2.1. Trained action classification

We separately test two datasets and learned a lexicon (see Figure 3 and 4) for each dataset used for action recognition. Comparing the learned lexicon with the ground truth lexicon, one finds the lexicon for MSRC-12 and WorkoutSu-10 are almost consistent with ground truth except the semantic pose T_4 in MSRC-12 dataset. Notice also that corresponding pose of semantic pose T_4 is confused with T_5 . These two semantic poses are used by only one action “Push right”, which reduces the performance of the proposed method by fewer co-occurrence of semantic pose and visual pose candidates. The variation among subjects performing this action also results in confusion as subjects may ignore the elementary action from T_1 to T_4 .

The learned lexicon was further verified through action recognition. The accuracy gained with MRSC-12 and WorkoutSu-10 dataset are respectively 85.86% and 98.71%. Comparative results with discriminative models are shown in Table 1. Although the performance on MSRC-12 is slightly worse than discriminative models [26, 12, 27], it is a reasonable result when considering the fact that we do not model reference object (G) and simply consider the whole body as object. Consequently, it is hard to distinguish actions “Googles and “Had enough which have particular reference objects. The result based on WorkoutSu-10 dataset outperforms the discriminative model RDF [12] even though the number of instances in [12] was 300 fewer than in our experiment.

To demonstrate the performance of semantic action recognition with the aid of textual instruction, we compare the proposed method with state-of-the-art semantic learning methods. As attributes [19, 20] are hard to describe and not comparable, we opt to compare it with state-of-the-art latent semantic learning. The comparisons are made with generative

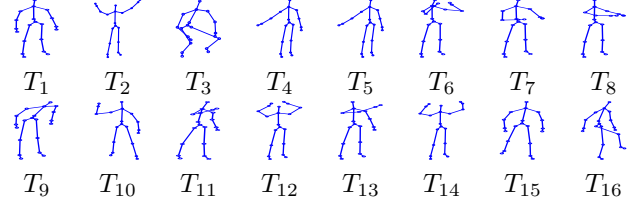


Fig. 3: A lexicon based on MSRC-12 dataset.

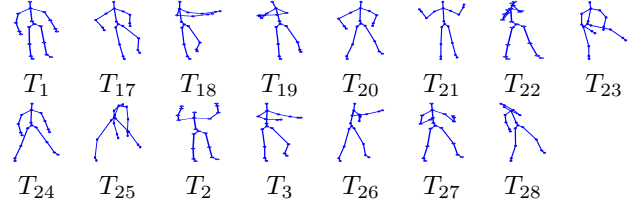


Fig. 4: A lexicon based WorkoutSu-10 dataset.

Table 1: Comparative results: trained action classification.

Methods	MSRC-12	WorkoutSu-10
Cov3DJ [26]	91.70%	-
RDF [12]	94.03%	98%
ELC-KSVD [27]	90.22%	-
LDA[14]	74.81%	92.27%
HGM [18]	66.25%	82.37%
Ours	85.86%	98.71%

models and results are shown in Table 1; the proposed method can be categorised as a generative model. In particular, we compare it with classical latent Dirichlet allocation (LDA) model [14] and a hierarchical generative model (HGM) [18] which is a two-layer LDA. In both models, word is similar to visual pose candidate of the proposed method and topic is similar to semantic pose. Hence, in LDA the number of topics is equal to the number of semantic poses. In HGM, the number of global topics is same as the number of semantic poses. Comparative results show that the proposed method outperforms HGM and LDA, clearly indicating its semantic representation power for action recognition. Moreover, experiments investigating the role of k (set size of visual pose candidate) showed an increased rate for the proposed method with increased value of k , levelling off when k is about 5 or 6 times as many as the size of semantic pose set ².

4.2.2. Zero-shot classification

Zero-shot classification is to recognize an action that has not been trained before. Notice that the four semantic poses are shared in different actions among two datasets including T_1 , T_2 , T_3 and T_7 . This motivated the selection of all actions that used these semantic poses for experiments. We also con-

² Graphical result is available in the supplementary material.

Table 2: Results of zero-shot classification.

Testing set		Accuracy
Single action	Lift arms	92.72%
	Duck	61.40%
	Wind up	54.70%
Composite activity	A_1 then A_2	99.30%
	B_2 then B_3	100%
	C_1 then C_2	98.67%

structured a synthesized set of composite activities by concatenating single actions from WorkoutSu-10 in order to enlarge the number of test actions. Specifically, training actions include: *Cheap weapon*, *Beat both*, *Hip flexion* (A_1), *Trunk rotation* (A_2), *Hip adductor stretch* (B_2), *Hip adductor stretch* (B_3), *Curl-to-press* (C_1) and *Squats* (C_2). Three single actions and three composite activities are used for testing and recognition accuracy is shown in Table 2. Results demonstrate that the proposed method based on action semantic poses is effective in zero-shot action recognition.

5. CONCLUSION

This paper has presented a novel method of learning a pose lexicon that consists of semantic poses and visual poses. Experimental results showed that the proposed method can effectively learn the mapping between semantic and visual poses and was verified in both pre-trained and zero-shot action recognition.

The proposed method can be easily extended to semantic action recognition based on RGB or depth datasets and provides a foundation to build action-verb and activity-phrase hierarchies. A unique large lexicon can also be learned for action recognition involving different datasets. In addition, the future work about representing semantic poses will be explored to improve the proposed method through modelling reference objects and middle part of trajectories since we only considered start and end positions of trajectories.

6. REFERENCES

- [1] W. Li, Z. Zhang, and Z. Liu, "Expandable data-driven graphical modeling of human actions based on salient postures," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1499–1510, 2008.
- [2] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Comput. Vision Image Underst.*, vol. 115, no. 2, pp. 224–241, 2011.
- [3] P. Wang, W. Li, P. Ogunbona, Z. Gao, and H. Zhang, "Mining mid-level features for action recognition based on effective skeleton representation," in *Proc. Int. Conf. Digital Image Computing Techniques Appl.*, 2014.
- [4] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang, and P. Ogunbona, "Convnets-based action recognition from depth maps through virtual cameras and pseudocoloring," in *Proc. Annual Conf. ACM Multimedia*, 2015.
- [5] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Trans. Human-Mach. Syst.*, Accepted 2 November 2015.
- [6] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proc. Comput. Vision Pattern Recognit. Workshops*, 2010.
- [7] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *Proc. Comput. Vision Pattern Recognit.*, 2013.
- [8] A. Eweiri, M. S. Cheema, C. Bauckhage, and J. Gall, "Efficient pose-based action recognition," in *Proc. Asian Conf. Comput. Vision*. Springer, 2014, pp. 428–443.
- [9] L. Talmy, *Toward a cognitive semantics*, M. I. of Technology, Ed. MIT press, 2003, vol. 1.
- [10] S. Petrov, L. Barrett, R. Thibaux, and D. Klein, "Learning accurate, compact, and interpretable tree annotation," in *Proc. Annual Meeting Assoc. Comput. Ling.*, 2006.
- [11] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proc. SIGCHI Conf. Human Factor Comput. Syst.*, 2012.
- [12] F. Negin, F. Özdemir, C. B. Akgül, K. A. Yüksel, and A. Erçil, "A decision forest based feature selection framework for action recognition from RGB-Depth cameras," in *Image Anal. and Recognit.* Springer, 2013, pp. 648–657.
- [13] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. Comput. Vision Pattern Recognit.*, 2008.
- [14] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vision*, vol. 79, no. 3, pp. 299–318, 2008.
- [15] Y. Wang and G. Mori, "Human action recognition by semi-latent topic models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1762–1774, 2009.
- [16] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learning*, vol. 42, no. 1-2, pp. 177–196, 2001.
- [17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
- [18] S. Yang, C. Yuan, W. Hu, and X. Ding, "A hierarchical model based on latent dirichlet allocation for action recognition," in *Proc. Int. Conf. Pattern Recognit.*, 2014.
- [19] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Proc. Comput. Vision Pattern Recognit.*, 2011.
- [20] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Attribute regularization based human action recognition," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 10, pp. 1600–1609, 2013.
- [21] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele, "Script data for attribute-based recognition of composite activities," in *Proc. European Conf. Comput. Vision*, 2012.
- [22] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko, "Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proc. Int. Conf. Comput. Vision*, 2013.
- [23] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *Proc. European Conf. Comput. Vision*, 2002.
- [24] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection," in *Proc. Int. Conf. Comput. Vision*, 2013.
- [25] P. Koehn, *Statistical machine translation*. Cambridge University Press, 2009.
- [26] M. E. Hussein, M. Torki, M. A. Gawayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptor on 3D joint locations," in *Proc. Int. Joint Conf. Artificial Intelligence*, 2013.
- [27] L. Zhou, W. Li, Y. Zhang, P. Ogunbona, D. T. Nguyen, and H. Zhang, "Discriminative key pose extraction using extended LC-KSVD for action recognition," in *Proc. Int. Conf. Digital Image Computing Techniques Appl.*, 2014.

Supplementary material:

In this supplementary material, the textual instructions of the actions in the MSRC-12 gesture [11] and WorkoutSu-10 exercise [12] datasets are described in the first section. The second section details the extracted semantic poses of the two datasets. In the third section, the experimental results on the recognition accuracy vs. the number of visual pose candidates are presented.

Textual instructions

Beginning with a starting configuration of the body, the instructions consist of a sequence description of the elementary motions that constitute an action. The instructions can be understood and followed by a subject with average literacy to perform the action.

MSRC-12

- Lift outstretched arms:
 - Begin in standing position with arms beside body.
 - Raise outstretched arms from sides to shoulder height until overhead.
 - Return to starting position.
- Duck:
 - Begin in standing position with arms beside body.
 - Slightly bend knee.
 - Return to starting position.
- Push right:
 - Begin in standing position with arms beside body.
 - Bring right hand in front of tummy.
 - Slide right hand towards the right until same plane as torso.
 - Return to starting position.
- Goggles:
 - Begin in standing position with arms beside body.
 - Raise hands through tummy, chest until eyes.
 - Return to starting position.
- Wind up:
 - Begin in standing position with arms beside body.
 - Raise arms through tummy, chest until shoulder height.
 - Bring arms back and return to starting position.
- Shoot:
 - Begin in standing position with arms beside body.
 - Stretch arms out in front of body with clasped hands.
 - Slightly bring up hands.
 - Return to starting position.
- Bow:
 - Begin in standing position with arms beside body.
 - Bend forwards with torso parallel with the floor.
 - Return to starting position.
- Throw:
 - Begin with standing position.
 - Step left foot forward while raising right hand overhead and back to the point of maximum external shoulder rotation.
 - Accelerate right arm forward until elbow is completely straight.
 - Return to starting position.
- Had enough:
 - Begin with standing position.
 - Raise arms with hands on the head.
 - Return to starting position.
- Change weapon:
 - Begin with standing position.
 - Reach over left shoulder with right hand.
 - Bring both hands in front of tummy as if they are holding something.
 - Return to starting position.
- Beat both:
 - Begin with standing position.
 - Raise hands to the head level.
 - Beat hands sideways for a few times.
 - Return to starting position.
- Kick:
 - Begin with standing position.
 - Step left foot forward while raising right foot to the right knee.
 - Straighten right leg forward.
 - Return to starting position.

WorkoutSu-10

- Hip flexion (A_1)
 - Begin in standing position with arms beside body.
 - Place hands on the waist.
 - Flex left leg at the hip up to 90 degree and bend the knee.
 - Return to starting position.
- Torso rotation (A_2)
 - Begin in standing position with arms beside body.
 - Outstretch arms at shoulder height with clasped hands and flex left leg at knee.
 - Rotate torso towards left side at almost 45 degrees.
 - Rotate torso to the right.
 - Return to starting position.
- Lateral stepping (A_3)
 - Begin in standing position with arms beside body.
 - Place hands on the waist.
 - With right foot step laterally to the right.
 - Bring left foot beside right foot.
 - Repeat the last two elementary actions for several times.
 - With left foot step laterally to the left.
 - Bring right foot beside left foot.
 - Repeat the last two elementary actions for several times.
 - Return to starting position.
- Thoracic rotation (B_1)
 - Begin in standing position with arms are flexed at elbow and hands raised at shoulder height.
 - Rotate torso to left side.
 - Return to starting position.
- Hip Adductor stretch (B_2)
 - Begin in standing position with arms beside body.
 - Spread legs beyond shoulder width.
 - Shift body weight to right leg by bending the knee up to 90 degree and straighten left leg.
 - Return to starting position.
- Hip adductor stretch (B_3)
 - Begin in standing position with arms beside body.

- Spread legs beyond shoulder width.
- Bend forwards with torso parallel with the floor while arms are hanging down.
- Return to starting position.

- Curl-to-press (C_1)
 - Begin in standing position with arms beside body.
 - Flex arms at elbow and raise arms in front of body and overhead.
 - Return to starting position.
- Freestanding squats (C_2)
 - Begin in standing position with arms beside body.
 - Bend knee up to 90 degrees and arms are outstretched at shoulder height in front of the body.
 - Return to starting position.
- Transverse horizontal punch (C_3)
 - Begin in standing position with arms beside body.
 - Rotate torso towards left side and punch with right hand in front of the body until arm is straight.
 - Rotate torso towards right side and punch with left hand in front of the body until arm is straight.
 - Return to starting position.
- Oblique stretch (C_4)
 - Begin in standing position with arms beside body.
 - Spread legs beyond shoulder width.
 - Raise left arm laterally overhead and lean torso towards the right.
 - Return to starting position.

Semantic poses

Semantic poses extracted from the MSRC-12 gesture and WorkoutSu-10 exercise datasets are show in Table 3 and 4 and their linguistic descriptions are as follows.

T_1 Arms are beside legs.

T_2 Arms are overhead with elbows above shoulder level.

T_3 Thighs at an angle to the shins.

T_4 Right arm is in front of stomach.

T_5 Horizontally outstretched right arm is in the right of the body.

T_6 Hands are on the corresponding eyes.

- T_7 Two arms are in front of stomach.
- T_8 Horizontally outstretched arms are in front of the body.
- T_9 Torso is parallel to floor.
- T_{10} Left foot is one step in front of the right foot. Right elbow is at shoulder level.
- T_{11} Left foot is one step front of the right foot. Horizontally outstretched right arm is in front of the body.
- T_{12} Hands are on the head.
- T_{13} Horizontally outstretched right arm is in front of the body.
- T_{14} Two elbows are at shoulder level.
- T_{15} Left foot is one step in front of the right foot.
- T_{16} Raised and outstretched right leg is in front of the body.
- T_{17} Forward raised left knee is with thigh at almost 90 degrees to the shin.
- T_{18} The trunk faces the left side at almost 45 degrees. Horizontally outstretched arms are in front of the body. Left shin is at almost same level as right knee.
- T_{19} The trunk faces the right side at almost 45 degrees. Horizontally outstretched arms are in front of the body. Right shin is at almost same level as left knee.
- T_{20} Two hands on the hips. Thighs at a slight angle to the shins.
- T_{21} Raised two arms are beside the body with hands at almost shoulder level.
- T_{22} The trunk faces one side at almost 45 degrees. Raised two arms are beside the body with hands at almost shoulder level.
- T_{23} One thigh at an angle to the corresponding shin. The other leg is outstretched.
- T_{24} Separate legs are much wider than shoulder.
- T_{25} Separate legs are much wider than shoulder. Torso is parallel to floor and hanging arms are down besides legs.
- T_{26} Trunk faces the left side at almost 90 degrees. Horizontally outstretched right arm is in front of the body.
- T_{27} Trunk faces the right side at almost 90 degrees. Horizontally outstretched left arm is in front of the body.
- T_{28} The trunk is bent obliquely to one side. The stretched opposite side arm is over the head.

Table 3: Semantic poses used in MSRC-12 dataset.

Gestures	Semantic poses
Lift outstretched arms	T_1, T_2
Duck	T_1, T_3
Push right	T_1, T_4, T_5
Goggles	T_1, T_6
Wind up	T_1, T_7, T_2, T_7
Shoot	T_1, T_8
Bow	T_1, T_9
Throw	T_1, T_{10}, T_{11}
Had enough	T_1, T_{12}
Change weapon	T_1, T_{13}, T_7
Beat both	T_1, T_{14}, T_2
Kick	T_1, T_{15}, T_{16}

Table 4: Semantic poses used in WorkoutSu-10 dataset.

Exercises	Semantic poses
Hip flexion (A_1)	T_1, T_{17}
Trunk rotation (A_2)	T_1, T_{18}, T_{19}
Lateral stepping (A_3)	T_1, T_{20}
Thoracic rotation (B_1)	T_{21}, T_{22}
Hip adductor stretch (B_2)	T_1, T_{23}
Hip stretch (B_3)	T_1, T_{24}, T_{25}
Curl-to-press (C_1)	T_1, T_2
Freestanding squats (C_2)	T_1, T_3
Transverse horizontal punch (C_3)	T_1, T_{26}, T_{27}
Oblique stretch (C_4)	T_1, T_{28}

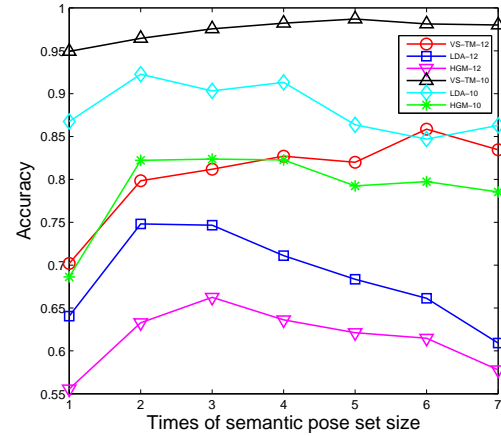


Fig. 5: Recognition accuracy versus the number of visual pose candidates.

Impact of the number of visual pose candidates

The recognition accuracy versus the number of visual pose candidates is shown in Figure 5, where 12 and 10 in the

legends represent MSRC-12 and WorkoutSu-10 datasets respectively. Specifically, the number of visual pose candidates ranges from one to seven times as many as the number of semantic poses. Results have shows that the recognition rate of the propose method increases first with the increase of the number of visual pose candidates and then levels off when the number of visual pose candidates reaches 5 or 6 times as many as the number of semantic poses.

We also compare the proposed method with the latent Dirichlet allocation (LDA) model [14] and a hierarchical generative model (HGM) [18] on the MSRC-12 and WorkoutSu-10 datasets, where the number of words was set to one to seven times as many as the number of topics in LDA and the number of global topics in HGM respectively. Experimental results have shown that the proposed method is more robust than LDA and HGM with respect to the number of visual pose candidates.